Received: 7 March 2019 | Revised: 24 July 2019 | Accepted: 26 July 2019 | Published on: 30 October 2019

DOI: 10.1002/qj.3638

RESEARCH ARTICLE

Quarterly Journal of the Royal Meteorological Society RMetS

# Regime-dependent statistical post-processing of ensemble forecasts

Sam Allen | Christopher A. T. Ferro | Frank Kwasniok

Department of Mathematics, University of Exeter, Exeter, UK

**Correspondence**
Sam Allen, University of Exeter, Laver Building, North Park Road, Exeter, EX4 4QE, UK.
Email: sa495@exeter.ac.uk

**Funding information**
Natural Environment Research Council, NE/N008693/1

**Abstract**

A number of realizations of one or more numerical weather prediction (NWP) models, initialised at a variety of initial conditions, compose an ensemble forecast. These forecasts exhibit systematic errors and biases that can be corrected by statistical post-processing. Post-processing yields calibrated forecasts by analysing the statistical relationship between historical forecasts and their corresponding observations. This article aims to extend post-processing methodology to incorporate atmospheric circulation. The circulation, or flow, is largely responsible for the weather that we experience and it is hypothesized here that relationships between the NWP model and the atmosphere depend upon the prevailing flow. Numerous studies have focussed on the tendency of this flow to reduce to a set of recognisable arrangements, known as regimes, which recur and persist at fixed geographical locations. This dynamical phenomenon allows the circulation to be categorized into a small number of regime states. In a highly idealized model of the atmosphere, the Lorenz '96 system, ensemble forecasts are subjected to well-known post-processing techniques conditional on the system's underlying regime. Two different variables, one of the state variables and one related to the energy of the system, are forecasted and considerable improvements in forecast skill upon standard post-processing are seen when the distribution of the predictand varies depending on the regime. Advantages of this approach and its inherent challenges are discussed, along with potential extensions for operational forecasters.

**KEYWORDS**

ensemble prediction, forecast guidance, probabilistic weather forecasting, recalibration, statistical post-processing, weather regimes

## 1 | INTRODUCTION

The atmosphere is a chaotic dynamical system. Hence, weather forecasts are heavily reliant on a perfect measure of their initial conditions, something that is never achieved in practice. To address this, dynamical numerical weather prediction (NWP) models are run from a variety of initial conditions to obtain a sample of distinct forecasts (Leith, 1974). In addition to error in the initial state, the models themselves are imperfect. The result is a biased, typically underdispersed

and thus overconfident ensemble forecast. To calibrate the ensemble forecasts, they are often subjected to statistical post-processing. These statistical methods serve as a way of issuing a well-calibrated probabilistic forecast in observation space given NWP realizations in the model's phase-space (Stephenson *et al.*, 2005).

Statistical post-processing removes systematic errors present in the NWP models by detecting and correcting relationships between past forecasts and the resulting observations. However, these relationships are not necessarily stationary. Hamill *et al.* (2017) remarks that biases in NWP output may vary with season, spatial location and other factors that systematically influence the model error. That is, the relationship between the NWP model and the atmosphere may change under different circumstances. If such circumstances can be identified then it may be possible to incorporate this additional information into established post-processing methods.

We postulate that the relationship between the NWP model and the atmosphere changes depending on the concurrent behaviour of the atmospheric circulation. This nonlinear, chaotic flow falls into recognisable, large-scale structures called regimes, regarded as metastable equilibria of the flow's phase-space. These atmospheric, or weather, regimes characterize the low-frequency variability of the circulation, which at these equilibria exhibits noticeably regimented behaviour; the flow patterns persist and recur at fixed geographical locations. Examples of this dynamical behaviour include persistent anomalies in geopotential height fields (Dole and Gordon, 1983), such as blocking, and teleconnection patterns – highly negatively correlated variables situated at widely separated spatial locations (Wallace and Gutzler, 1981).

There exist ample studies exploring the nature of low-frequency variability in the atmosphere and this article provides only a basic introduction to the regime paradigm, highlighting relevant results and focussing on their statistical representation. For a considerably more thorough review of the extant literature, readers are diverted to Hannachi *et al.* (2017) and references therein. The fundamental concept is that the atmospheric circulation can be decomposed into just a few metastable equilibrium states and that transitions between these regimes can thus be used to describe the continuous evolution of the atmosphere (Franzke *et al.*, 2011).

The circulation is primarily responsible for the weather that we experience and further justification for its inclusion in statistical post-processing can be found in past literature. Robertson and Ghil (1999) conclude that weather regimes affect the frequency and magnitude of temperature and precipitation events, while Neal *et al.* (2016) proposes that more extreme weather events have a higher probability of occurring in certain circulation types, suggesting the predictability of the atmosphere may vary for different regimes.

Messner *et al.* (2017) highlights the potential improvements to post-processing when a variety of atmospheric variables are included in the statistical models, rather than relying solely on the forecasts issued by the ensemble members. Weather regimes implicitly incorporate the behaviour of other atmospheric variables without suffering from challenges such as overfitting and variable selection that are induced by using a large number of possible predictors.

Perhaps the most promising reason for believing that there exists a different relationship between the model and the atmosphere in different weather regimes can be found in Ferranti *et al.* (2015). The paper assesses the performance of raw ensemble forecasts when the atmosphere resides in four atmospheric regimes – the positive and negative phases of the North Atlantic Oscillation (NAO), an Atlantic Ridge and European Blocking – concluding that the skill of medium-range weather forecasts changes when initialised in certain regimes.

The remainder of the article is organized as follows. A discussion of the general problem and the choice of methodology to investigate is provided in section 2. Section 3 introduces a highly idealized model of the atmosphere, the Lorenz (1996) system, in which this methodology will be tested. The post-processing methods and forecast verification techniques are presented in section 4, with the corresponding results for the simulation study displayed in section 5. Section 6 discusses the practicalities of the method and extensions to operational forecasts, while also concluding.

## 2 | METHODOLOGY

The focus of this work is to extend current methods of statistically post-processing ensembles of weather forecasts, which generate forecasts in the form of predictive probability distributions, $G(v|\boldsymbol{f}, \boldsymbol{\theta})$. Here, $\boldsymbol{f} = (f_1, f_2, \ldots, f_M)$ is an ensemble forecast comprising $M$ ensemble members, $\boldsymbol{\theta}$ is a vector of parameters and $G$ is a parametric distribution chosen for the weather variable of interest. We consider the case where the response, or verification, $v$ is univariate; however, the method could easily be extended for multivariate post-processing.

Although there has been some debate on the irrefutable presence of weather regimes, they are a useful feature in this framework. Defining regimes to exhibit persistence renders the time spent transitioning between states negligible compared to time spent in the regimes. The regime states thus form a mutually exclusive, collectively exhaustive (MECE) partition of the atmosphere's phase-space.

This provides a helpful reduction but it is possible to proceed without it. If the circulation can be quantified by some continuous metric, $\rho$, then the predictive distributions could simply be extended to include this metric as an additional

variable in the recalibration:

$$G(v \mid \boldsymbol{f}, \boldsymbol{\theta}, \rho).$$

Using a continuous measure allows the flow to be represented on a spectrum and, rather than harshly binning the circulation into a finite number of regimes, it permits a degree of membership to several states to be quantified. In reality, although indices exist that measure how much the atmosphere resembles commonly recognized weather regimes such as the NAO, the Arctic Oscillation and the Pacific–North American pattern, there is no recognized method of objectively condensing the flow over some spatial domain into a single continuous metric.

Suppose instead that a finite number, $R$, of regimes in the atmosphere are identified. If an underlying regime can accurately be attributed to a forecast, then recalibration can be performed conditional on this atmospheric state. For example, when forecasting in each regime, the post-processing methods could use a separate set of model parameters:

$$G(v \mid \boldsymbol{f}, \boldsymbol{\theta}_r),$$

or even specify a distinct distribution:

$$G_r(v \mid \boldsymbol{f}, \boldsymbol{\theta}_r),$$

for $r = 1, \ldots, R$.

More generally, the forecast distribution can be written as a mixture of predictive distributions that depend on the regime:

$$G(v \mid \boldsymbol{f}, \boldsymbol{\theta}) = \sum_{r=1}^{R} w_r G_r(v \mid \boldsymbol{f}, \boldsymbol{\theta}_r),$$

where the weight $w_r$ represents the probability of the atmosphere residing in regime $r$, allowing the model to account for uncertainty present when attributing the forecast to a regime.

This article focusses on these regime-based extensions; the idea of introducing a continuous metric to measure circulation is not investigated. Discretizing the flow like this places fewer restrictions on any model parameters, allowing for more flexibility in the statistical recalibration models.

Although we focus here on weather regimes, this approach is suitable for any grouping of the forecasts in which different model biases might be expected. Similar extensions to statistical post-processing have been implemented previously in the hope of attaining more skilful forecasts of extreme wind-speed events. Lerch and Thorarinsdottir (2013) and Baran and Lerch (2015) apply a regime-switching approach that issues a separate predictive distribution depending on whether or not the ensemble median lies above some threshold, suggesting that biases in the forecasts depend on the predicted values themselves.

Rather than using a fixed threshold, Baran and Lerch (2016) extend this idea further by utilizing a mixture of the predictive distributions, with weight parameters that are estimated simultaneously with the coefficients of the component distributions. Although the regime-switching approaches implicitly assume that biases differ between two or more distinct configurations of the atmosphere, they do not necessarily refer to weather regimes. Gneiting *et al.* (2006), however, finds that skilful short-range forecasts of wind speed are obtained when separate statistical models are fitted depending on the local prevailing wind direction.

Statistical post-processing corrects systematic errors in the raw ensemble by exploiting relationships between archived forecasts and their corresponding verifications. Thus, a training dataset of historical forecasts and observations – forecast–observation pairs – is required, from which relationships can be identified and parameters can be estimated. Continual adjustments to NWP models often limit the training data available to operational forecasters. The flow, on the other hand, is a product of the atmosphere only and is not dependent on the forecast. Therefore the regimes need not be estimated from the training data, they can be discerned from a much larger set of observations.

However, regimes are hidden and must be inferred from other, observable, variables. This can be circumvented by converting these dynamical phenomena to statistical artefacts. A variety of statistical approaches have been used to detect atmospheric equilibria including pattern correlation analysis (Horel, 1985), probability density analysis (Kimoto and Ghil, 1993), clustering algorithms (Cheng and Wallace, 1993; Smyth *et al.*, 1999; Kondrashov *et al.*, 2004) and hidden Markov models (Majda *et al.*, 2006). Unfortunately, the regimes identified are not always robust to the method used; a number of these studies have considered wintertime geopotential height anomalies in the Northern Hemisphere, yet have yielded contrasting regime-like behaviour.

There has been extensive work on regime detection and this framework assumes only that the statistical representations are reasonable approximations of their dynamical counterparts – beyond this, the choice among methods is arbitrary. The regimes are hereafter assumed to be known.

The regime-dependent approaches rely on the ascribing of forecasts to an underlying state. Thus, a method is required that condenses information regarding the atmosphere into just one of a number of predetermined regimes. Since each NWP forecast represents a simulated trajectory of the atmosphere, this method should also be able to predict a regime given the NWP output. Therefore, provided forecasts are of the same spatial scale as the regimes, each ensemble member provides an estimate of the atmospheric state; members can be matched with the regime that is statistically the closest (Neal *et al.*, 2016).

If this method accurately assigns an ensemble forecast–observation pair to a regime then the training dataset, in acting as a sample of the system's phase-space, can be stratified into $R$ MECE subsets. Relationships can be identified between the ensemble forecasts and the observations from each of the separate subsets, including the estimation of a new set of parameters. However, the regime of a forecast is not unique; the underlying regime may change throughout the forecast and thus a time at which to define the regime must be chosen.

We seek the time at which the disparities between the model–atmosphere relationships are largest. There are two intuitive options: the state of the atmosphere at the forecast's initialisation time or at its validation time. In order to exploit past regime-dependent relationships, the regime of a new forecast should be defined in the same way as those in the training data. If the regime is defined at the initialisation time then it can be deduced (or estimated at least) from observed data, and thus does not rely on the NWP model being able to capture the regime structure present in the atmosphere.

In this case, the training data can be stratified into subsets depending on the regime of the forecast at its initialisation time and separate post-processing parameters can be estimated for each subset. Any new forecast would then be assigned to a regime in the same way and post-processed using the parameters estimated from the corresponding training subset. This assumes that all ensemble members estimate the same regime at the initialisation time and that a small perturbation to the analysis is not sufficient to alter the large-scale state of the atmosphere.

However, since the length of a medium-range weather forecast may exceed the average duration of a weather regime, the atmospheric state will often change throughout the forecasting period. Therefore, conditioning on the regime at the initialisation time may result in losing some information regarding the occurrence of different weather events in different regimes, contradicting some of the reasons for believing this method may be successful, such as extreme events occurring more frequently in certain regimes.

Using the regime of the atmosphere at the forecast's validation time does not suffer from these problems and therefore may be expected to yield more heterogeneous relationships between the model and the atmosphere. However, the regime at the validation time is not known and hence the forecast could not be assigned to exactly one regime in the same way that those in the training data were.

The regime of a forecast at its validation time could instead be estimated using the regimes approximated from the ensemble members. This yields $M$ regime estimates for each ensemble forecast and a sensible approach might be to use the proportion of ensemble members predicting a regime as the probability of the atmosphere residing in that regime. From this, ensembles could, for example, be calibrated using a mixture of post-processing models with corresponding weights. Here, since every forecast–observation pair would not necessarily be assigned to exactly one regime, rather than stratifying the training data into subsets for each regime and estimating a separate set of parameters from each subset, a model averaging technique could be applied in which all parameters are estimated simultaneously.

This extension would be particularly well-suited to methods such as member-by-member post-processing (Van Schaeybroeck and Vannitsem, 2015) which corrects each ensemble member individually to yield forecasts in the form of a calibrated ensemble rather than a predictive distribution. In this setting, each ensemble member produces an estimate of a regime and so could be post-processed conditional on its own regime prediction.

In reality, it would be possible to use the state of the atmosphere at any intermediate time of the forecast, or even at any time prior to forecasting if such information were available, but these are yet more sensitive to the assumptions and challenges described above.

Section 4 reintroduces Non-homogeneous Gaussian Regression (NGR), also commonly referred to as Ensemble Model Output Statistics (EMOS), and Bayesian Model Averaging (BMA), and offers examples of possible extensions to these familiar statistical post-processing methods using the regime paradigm. A separate extension is considered when defining the regime at the initialisation time and at the validation time.

## 3 | LORENZ '96 SYSTEM

The methodology described in the previous section is implemented in a highly idealized model of the atmosphere, the Lorenz (1996) system. Its chaotic nature lends itself to simulations of weather forecasts and the trialling of statistical post-processing methods (Roulston and Smith, 2003; Wilks, 2006; Williams *et al.*, 2014). A coupled system containing both larger-scale variables, $X_k$, and subgrid-scale variables $Y_{j,k}$ is used to emulate the atmosphere:

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - \frac{hc}{b}\sum_{j=1}^{J} Y_{j,k};$$
$$\frac{dY_{j,k}}{dt} = -cbY_{j+1,k}(Y_{j+2,k} - Y_{j-1,k}) - cY_{j,k} + \frac{hc}{b}X_k, \quad (1)$$

for $k = 1, \ldots, K$ and $j = 1, \ldots J$. The system exhibits cyclic boundary conditions, $X_k = X_{k+K}$, $Y_{j,k} = Y_{j,k+K}$ and $Y_{j+J,k} = Y_{j,k+1}$.

The parameter values used are $K = 8$, $J = 32$, $F = 20$, $h = 1$, $b = 10$ and $c = 10$, and the system is numerically integrated forward in time using a fourth-order Runge–Kutta scheme with a time step of $dt = 0.001$. Christensen *et al.* (2015) showed that with these parameters the system exhibits regime-like behaviour, transitioning between two

distinct states. The regimes are defined using a pre-specified diagnostic:

$$\sum_{k=1}^{\frac{K}{2}} \mathrm{cov}\left(X_k, X_{k+\frac{K}{2}}\right), \tag{2}$$

where $\mathrm{cov}(X_i, X_j)$ denotes the covariance between the $i$th and $j$th components of the vector of state variables $\boldsymbol{X}$, calculated over a time series of length one model time unit (MTU; corresponding to 5 days) directly preceding the time of interest. The system resides in regime A if this covariance diagnostic is positive and regime B if it is negative. As such, regime A is characterized by high amplitudes of wave-number 2, and regime B is dominated by wave-number 1.

Whereas in reality there is uncertainty regarding the regime, this diagnostic allows a regime to be known with certainty, and thus removes the need to account for any uncertainty regarding the state of the system.

The NWP model can be represented by equations that resolve only the large scales, since this is a common simplification of dynamical weather models:

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F. \tag{3}$$

In an effort analogous to improving the NWP model, this equation can be extended by including a quartic polynomial of the resolved variable, which acts as a kind of sub-grid model to account for the effect of the neglected variables $Y_{j,k}$:

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - (\beta_0 + \beta_1 X_k + \beta_2 X_k^2 + \beta_3 X_k^3 + \beta_4 X_k^4). \tag{4}$$

The parameters $\beta_0, \beta_1, \ldots, \beta_4$ are estimated by minimizing the mean squared difference between true and parametrized tendencies (Wilks, 2005; Kwasniok, 2012). The resulting coefficient estimates are shown in Table 1. This model is also numerically integrated through time using a fourth-order Runge–Kutta scheme, this time with a time step of $dt = 0.005$.

To trial the regime-dependent statistical post-processing approach, a training dataset is generated, comprising forecasts initialised at points 0.15 MTU apart, from which parameters are estimated. The resulting post-processing models are assessed using a test dataset, with forecasts initialised at intervals of 50 MTU akin to Wilks (2006). A fixed training dataset is used throughout, consisting of 20,000 forecast–observation pairs, and trajectories up to a lead time of 3 MTU (15 days) are considered. The statistical post-processing methods are evaluated over 50,000 ensemble forecasts and verifications.

**TABLE 1** Parameter estimates for the quartic polynomial in the NWP model

| Parameter | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ |
|---|---|---|---|---|---|
| Estimate | 0.209 | 1.45 | −0.0127 | −0.00728 | 0.000312 |

**TABLE 2** Average duration (MTU) of regimes A and B and the proportion of time the systems spend in each regime

| | Mean duration | | % of time | |
|---|---|---|---|---|
| | Reg. A | Reg. B | Reg. A | Reg. B |
| True system | 6.23 | 1.60 | 80 | 20 |
| NWP model | 12.13 | 1.61 | 88 | 12 |

Along each margin, $k$, ensembles are generated by adding a stochastic perturbation to the initialisation points, governed by a $N(0, 0.1^2)$ distribution, and integrating the NWP model through time starting at these perturbed points. Ensembles of size 20 are used throughout, though the results were found not to depend on the ensemble size. To allow for interchangeable members, these ensembles do not contain a control, or analysis, forecast.

There are now two different processes, the true system imitating the atmosphere (Equation 1) and a deterministic NWP model with which ensemble forecasts can be generated (Equation 4). Table 2 shows the average persistence time of the regimes, along with the corresponding proportion of time the system spends residing in each regime. In the true system, regime A persists for 6.23 MTU (31 days) on average, and regime B only 1.60 MTU (8 days). The NWP model captures the mean persistence time of regime B but severely overestimates the persistence of regime A. Therefore the model spends a larger proportion of time in this state than the atmosphere.

Two different quantities are to be predicted. The system is invariant under translation and hence all margins of $\boldsymbol{X}$ are statistically identical. Therefore, since we are interested in univariate post-processing approaches, only $X_1$ is considered. Secondly, the mean squared value of all $X_k$ variables is also forecasted. This quantity is labelled $E$ since it is proportional to the total energy of the system:

$$E = \frac{1}{K}\sum_{k=1}^{K} X_k^2. \tag{5}$$

To visualize the regime-like behaviour, Figure 1 shows a year-long time-period (73 MTU) of the predictands, $X_1$ and $E$, along with the covariance diagnostic and the corresponding regime. Large spells in regime A with intermittent periods in regime B reinforce the features displayed in Table 2. There is no obvious disparity in the behaviour of $X_1$ depending on the regime of the system and this is confirmed by a plot of the empirical distributions of the observations in Figure 2. $E$, on the other hand, does appear to vary with the regime, with lower values coinciding with the occurrence of regime B.

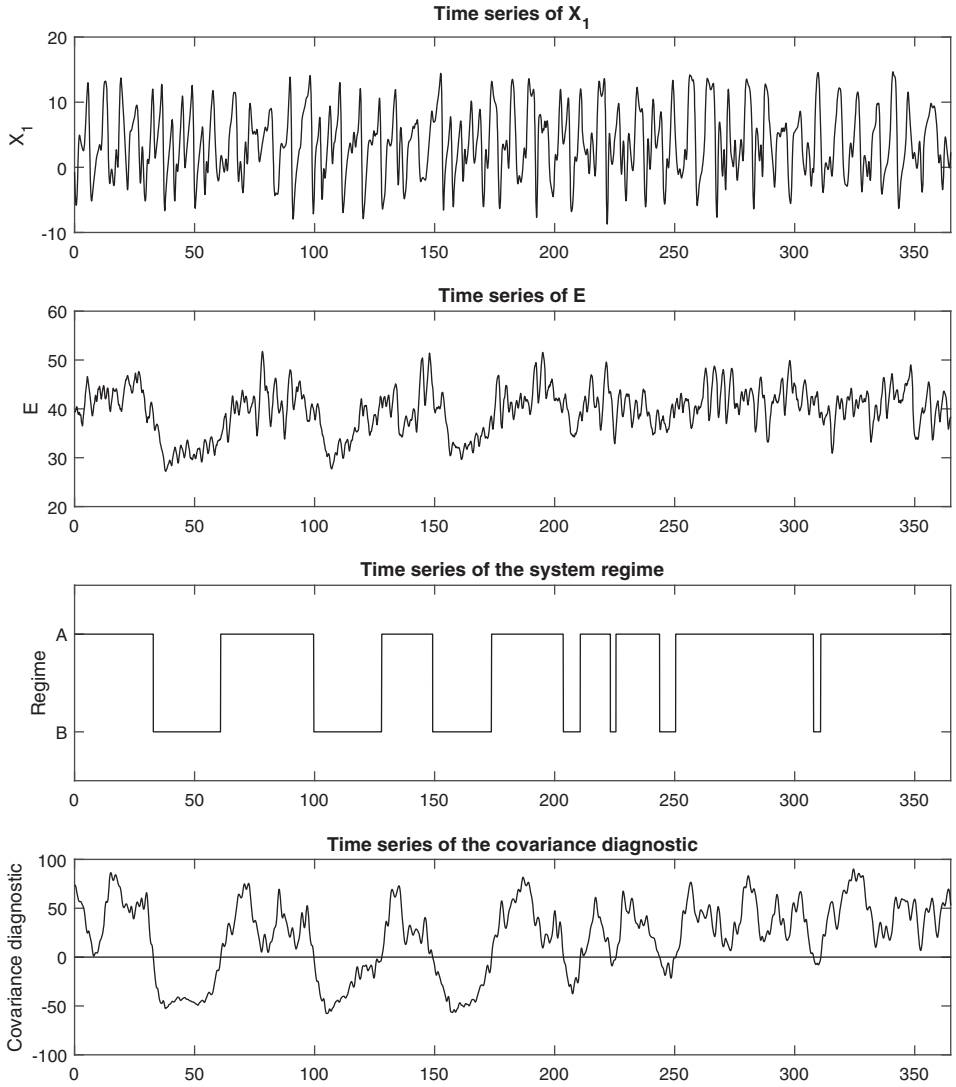


**FIGURE 1** Time series of the observed predictands, along with the concurrent regime and the associated value of the covariance diagnostic, for a year-long time period

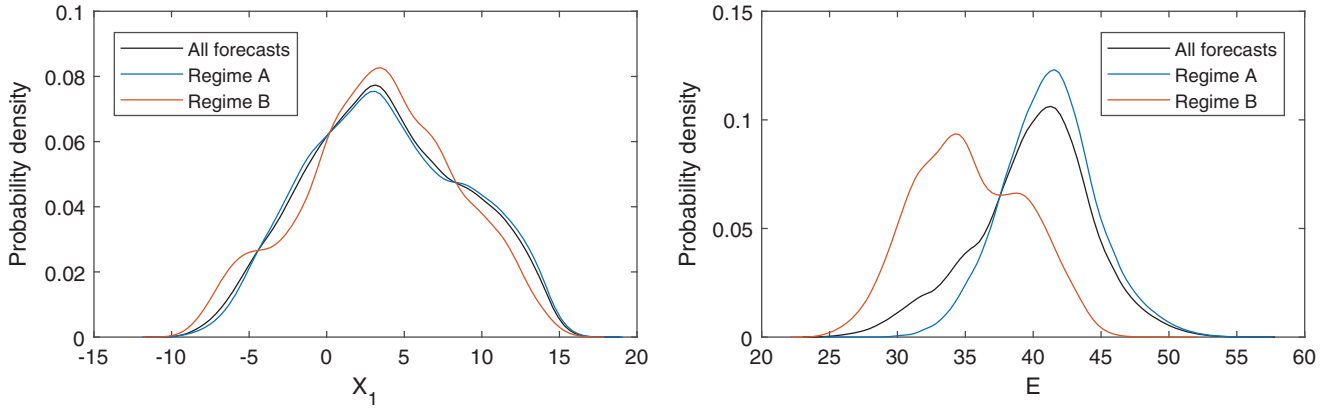


**FIGURE 2** Empirical distribution of $X_1$ (left) and $E$ (right) when the system resides in each regime

The similarity in the distributions of $X_1$ might appear disheartening since the method relies on discrepancies between the regimes; however, nothing can be deduced about the behaviour of the forecasts nor the predictability of the system in each regime. One particularly interesting attribute is that during prolonged spells in regime B the covariance diagnostic appears a lot less erratic, perhaps implying the system is more settled in this regime.

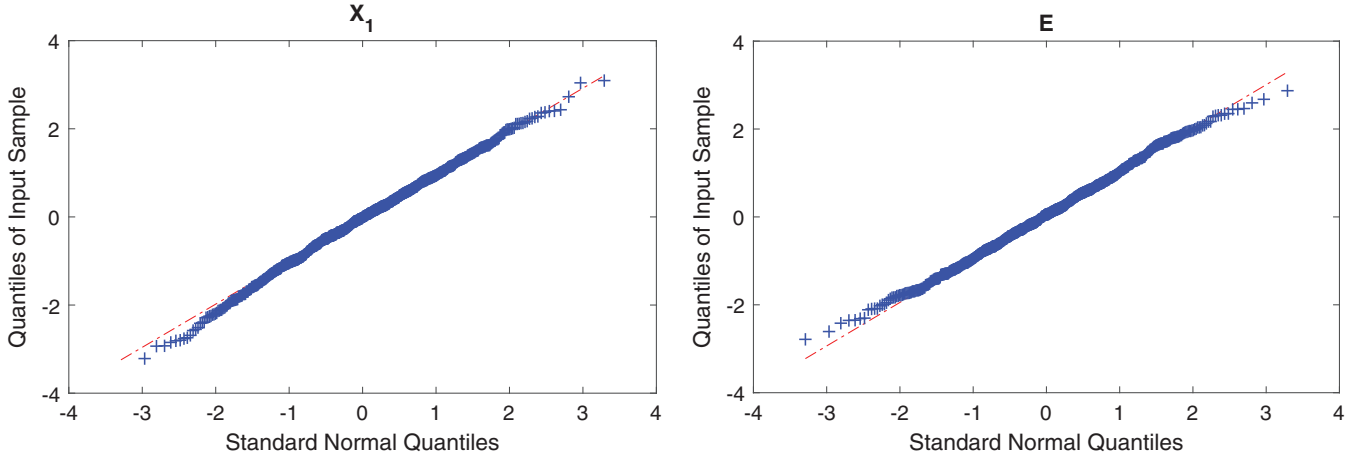


**FIGURE 3** Quantiles of a random sample of 1,000 standardized residuals from an NGR forecast plotted against the quantiles of a standard normal distribution. Shown when predicting $X_1$ (left) and $E$ (right) at a lead time of 3 days. Similar plots are found for forecasts from the two regimes, and also at other lead times

## 4 | STATISTICAL POST-PROCESSING

Consider a raw ensemble forecast $\boldsymbol{f}$ comprising $M$ members $f_1, f_2, \ldots, f_M$. Numerous techniques exist to statistically post-process the ensemble but we choose here to implement only the two most eminent methods, Bayesian Model Averaging (BMA) and Non-homogeneous Gaussian Regression (NGR). Whereas each ensemble member issues a point forecast – an instantaneous realization of phase-space – BMA and NGR generate probabilistic forecasts in the form of predictive distributions. These methods both assume that each verification, $v$, is a realization of a random variable, $V$, that follows a proposed statistical distribution conditional on the $M$ point predictions obtained from the raw ensemble members.

Despite deviations from Gaussianity in the marginal distributions of the observed values, suitable diagnostic checks, such as the quantile–quantile plots of the standardized residuals in Figure 3, show that the Normal distribution is an appropriate choice for the predictive distribution for both $V = X_1$ and $V = E$. $E$, by construction, is a positive quantity and using a Normal predictive distribution issues a non-zero probability of seeing a negative response. In this case, however, this probability is always negligibly small. A Gamma EMOS model was also implemented when forecasting $E$, but was found to perform worse than a Gaussian forecast distribution (not shown).

### 4.1 | Bayesian model averaging

BMA entails specifying a mixture of weighted component distributions that are centred around a linear adjustment of each ensemble member (Raftery *et al.*, 2005). Here we assume all members are interchangeable and hence equally weighted:

$$V|\boldsymbol{f} \sim \frac{1}{M}\sum_{m=1}^{M} N(\alpha + \beta f_m, \sigma^2). \tag{6}$$

The individual component distributions are Gaussian and the parameters $(\alpha, \beta, \sigma^2)$ are estimated by numerically maximizing the likelihood function or, equivalently, minimizing the negative log-likelihood (NLL) score. The NLL score for a mixture distribution with weights $w_m$ and Gaussian component distributions $N(\mu_m, \sigma_m^2)$ is

$$\text{nll} = -\log\left[\sum_{m=1}^{M} w_m \phi\left(\frac{v-\mu_m}{\sigma_m}\right)\right], \tag{7}$$

where $\phi(\cdot)$ is the standard Gaussian probability density function and $v$ is the corresponding observation. This score is then averaged over all $i = 1, \ldots, N$ forecasts in the training data to obtain the average NLL score, which in this case reduces to

$$\text{NLL} = -\frac{1}{N}\sum_{i=1}^{N}\log\left[\frac{1}{M}\sum_{m=1}^{M}\phi\left(\frac{v_i-\alpha-\beta f_{m,i}}{\sigma}\right)\right]. \tag{8}$$

In the regime paradigm we propose two different extensions to the model depending on when the regime of the forecast is defined. If the state of the atmosphere is defined at the intialisation time then the training data can be divided into subsets based upon the regime of the atmosphere at the forecast's initialisation time, and a separate set of parameters can be estimated for each regime $(\alpha_r, \beta_r, \sigma_r^2$ for $r = 1, \ldots, R)$ by minimizing the NLL score over each training subset:

$$\text{NLL} = -\frac{1}{N_r}\sum_{i=1}^{N_r}\log\left[\frac{1}{M}\sum_{m=1}^{M}\phi\left(\frac{v_i-\alpha_r-\beta_r f_{m,i}}{\sigma_r}\right)\right], \tag{9}$$

where $N_r$ is the number of forecast–observation pairs in the learning data defined to be in regime $r$. A new forecast could then simply be conditioned on one regime:

$$V|\boldsymbol{f}, r \sim \frac{1}{M}\sum_{m=1}^{M} N(\alpha_r + \beta_r f_m, \sigma_r^2). \tag{10}$$

This method is referred to as RDBMA-init.

Alternatively, if the regime is defined at the validation time then since BMA specifies a separate distribution around each ensemble member, every member can be post-processed conditional on its own regime prediction. Members corresponding to the same regime are assumed to be statistically indistinguishable and hence an extension of BMA to include groups of exchangeable ensemble members is implemented (Fraley *et al.*, 2010):

$$V|\boldsymbol{f} \sim \sum_{r=1}^{R} w_r \sum_{m=1}^{M_r} N(\alpha_r + \beta_r f_m, \sigma_r^2). \tag{11}$$

$M_r$ denotes the number of ensemble members that predict regime $r$, and $w_r$ is the probability of being in that regime at the validation time, with $\sum_{r=1}^{R} w_r = 1$. Fraley *et al.* (2010) estimates this probability using maximum-likelihood via the Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977), but here the groups of exchangeable ensemble members are determined by the outputs of the NWP model and hence are not known prior to forecasting. As a result, $M_r$ changes for each ensemble.

Using $w_r = \frac{M_r}{M}$ thus allows the probability to vary for each forecast, providing a more flexible estimate that was found to produce more skilful predictions.

In this case, forecast–observation pairs cannot be assigned to exactly one regime and therefore the parameters must be estimated simultaneously. This method is termed RDBMA-val and the corresponding objective function is

$$\mathrm{NLL} = -\frac{1}{N}\sum_{i=1}^{N} \log\left[\frac{1}{M}\sum_{r=1}^{R}\sum_{m=1}^{M_r} \phi\left(\frac{v_i - \alpha_r - \beta_r f_{m,i}}{\sigma_r}\right)\right]. \tag{12}$$

## 4.2 | Non-homogeneous Gaussian regression

Recognising the presence of a spread-skill relationship, Gneiting *et al.* (2005) introduced Non-homogeneous Gaussian Regression to extend the Normal linear regression model to include a variance which is dependent on the spread of the ensemble members. The mean and variance of the predictive distribution are linear functions of the ensemble mean, $\overline{f}$, and variance, $s^2$, respectively. The result is a heteroscedastic distribution of the form

$$V|\boldsymbol{f} \sim N(\alpha + \beta\overline{f}, \gamma + \delta s^2). \tag{13}$$

To estimate the parameters ($\alpha$, $\beta$, $\gamma$, $\delta$; with $\gamma$ and $\delta$ constrained to be positive) in the regression equations, the paper acknowledges that the coefficients should be those that minimize a proper score and therefore propose minimum continuous ranked probability score (CRPS) estimation. Gneiting *et al.* (2005) showed the CRPS for a forecast in the form of a Gaussian predictive distribution to be

$$\mathrm{crps}[N(\mu, \sigma^2), v] = \sigma\left\{\frac{v-\mu}{\sigma}\left[2\Phi\left(\frac{v-\mu}{\sigma}\right) - 1\right] + 2\phi\left(\frac{v-\mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}}\right\}, \tag{14}$$

where $\Phi(\cdot)$ is the standard Gaussian cumulative distribution function and $v$ again represents the observed value. The total CRPS is then the average of this score computed over all forecasts in the training data:

$$\mathrm{CRPS} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{crps}[N(\mu_i, \sigma_i^2), v_i]. \tag{15}$$

Similarly to BMA, if the regime is defined at the initialisation time then each forecast is in either regime A or regime B and the model (labelled RDNGR-init) becomes

$$V|\boldsymbol{f}, r \sim N(\alpha_r + \beta_r\overline{f}, \gamma_r + \delta_r s^2) \tag{16}$$

for $r = 1, \ldots, R$. Again, parameters are estimated by stratifying the training dataset using the regime of the atmosphere at the forecast's initialisation time and minimizing the CRPS separately for each training subset.

However, if the regime is defined at the forecast validation time then it cannot be determined with certainty and hence a probabilistic approach is applied. Let $p_r$ denote the proportion of ensemble members that predict regime $r$. Then a mixture model of $R$ separate distributions could be implemented, with weights determined by $p_r$. The predictive distribution is of the form

$$V|\boldsymbol{f} \sim \sum_{r=1}^{R} p_r N(\alpha_r + \beta_r\overline{f}, \gamma_r + \delta_r s^2). \tag{17}$$

This is essentially a model averaging technique that exploits the regime predictions of the ensemble members to calculate the model weights. The CRPS for a forecast in the form of a mixture distribution with $J$ Gaussian component distributions and weights $w_j$ is

$$\mathrm{crps}\left[\sum_{j=1}^{J} w_j N(\mu_{j,i}, \sigma_{j,i}^2), v_i\right] = \sum_{j=1}^{J} w_j A(v_i - \mu_{j,i}, \sigma_{j,i}^2)$$

$$-\frac{1}{2}\sum_{j=1}^{J}\sum_{k=1}^{J} w_j w_k A(\mu_{j,i}-\mu_{k,i}, \sigma_{j,i}^2+\sigma_{k,i}^2), \qquad (18)$$

where

$$A(\lambda,\xi^2) = 2\xi\phi\left(\frac{\lambda}{\xi}\right) + \lambda\left[2\Phi\left(\frac{\lambda}{\xi}\right)-1\right]$$

(Grimit *et al.*, 2006). For the conditional distribution in Equation 17 there is a component distribution for each regime. Therefore, in this case, $J$ in Equation 18 is equal to the number of regimes $R$, and the weights $w_j$ are given by $p_r$. This approach is referred to as RDNGR-val.

### 4.3 | Forecast verification

These statistical post-processing methods are applied to a sample of point forecasts to obtain a predictive distribution conditional on the ensemble output. Forecasters have come to seek predictive distributions that are sharp subject to being calibrated and both of these qualities can be assessed by verifying forecasts using proper scoring rules (Gneiting and Raftery, 2007). In the following section the CRPS is used to verify forecasts. NGR forecasts are assessed using the same loss function with which parameters were optimized in the training data, and Equation 18 can also be used to evaluate BMA forecasts.

Although this might appear to favour NGR since parameters are estimated using the same score that is used to verify the forecasts, similar results are obtained when using the NLL score to assess forecasts, and also when BMA parameters are optimized using minimum CRPS estimation.

These scores outline the overall forecast performance but concern lies more on the improvement gained from the new methodology than on the raw scores themselves. Therefore the continuous ranked probability skill-score (CRPSS) is also applied. Whereas skill-scores are typically implemented with a simple benchmark such as climatology, the reference forecast is taken here to be the equivalent forecast obtained via NGR or BMA at the same lead time. For example, if we let $H$ denote the predictive distribution obtained from regime-dependent post-processing, $G$ denote that obtained from standard post-processing and $v$ the corresponding observation, then for a proper score $S(\cdot,\cdot)$, the skill-score $S_s$ is

$$S_s = \frac{\langle S(G,v)\rangle - \langle S(H,v)\rangle}{\langle S(G,v)\rangle} = 1 - \frac{\langle S(H,v)\rangle}{\langle S(G,v)\rangle}, \qquad (19)$$

with $\langle\cdot\rangle$ denoting the average score over forecasts in the test dataset (Wilks, 2019). The skill-score can thus be interpreted as the percentage improvement in score upon current post-processing methods, gained from regime-dependent post-processing.

## 5 | RESULTS

Defining regimes using the covariance diagnostic (Equation 2) allows the regime of the system to be issued with certainty given that the observations 1 MTU preceding the time of interest are known. Therefore, characteristics of forecasts can be compared for those defined to be in each regime at the initialisation time.

The statistical properties of the forecasts indicate that there are in fact disparities in the forecast behaviour between the two regimes. Figure 4 shows that the ensemble variance, computed from forecasts in the training data, is much smaller on average when the system resides in regime B than in regime A. This is true when predicting $X_1$ or $E$, and is particularly apparent when the regime is defined at the initialisation time. Such differences in the variance suggest that weather events are more predictable, and that the ensemble forecasts suffer more from overconfidence, when in regime B.

Results are now shown separately when forecasting $X_1$ and $E$.

### 5.1 | Forecasting $X_1$

Ensemble forecasts assume that the ensemble members arise from the same generation mechanism as the observation and hence the rank of the verification when pooled with the ensemble members should be uniformly distributed. This assumption can be evaluated by using verification rank histograms to visualize the distribution of the ranks across all forecasts in the test data (Anderson, 1996; Hamill and Colucci, 1997; Talagrand, 1997). Rank histograms displayed in Figure 5 indicate that the raw forecasts are highly overconfident, with observations falling outside of the range of ensemble members for the majority of forecasts. This is yet more prevalent for those initialised in regime B.

Figure 5 also displays Probability Integral Transform (PIT) histograms for the predictive distributions issued by NGR and RDNGR-init for the same lead time. PIT histograms record the frequency with which values of the forecast cumulative distribution function, evaluated at the verification, $p = G(v)$, fall into a finite number of equally-sized bins. In order to ensure comparability between the rank and PIT histograms, 21 bins between 0 and 1 were chosen. Likewise, a uniform PIT histogram implies calibrated forecasts.

The PIT histograms show that post-processing the forecasts using NGR yields considerably more uniform histograms, and hence considerably better-calibrated forecasts, than the raw ensembles. However, Hamill (2001) demonstrates how uniform rank histograms can be obtained from a combination of poorly-calibrated forecasts, emphasising that the uniformity of rank histograms is a necessary but not sufficient condition for reliable predictions. In this case, the PIT histogram for forecasts in regime B becomes

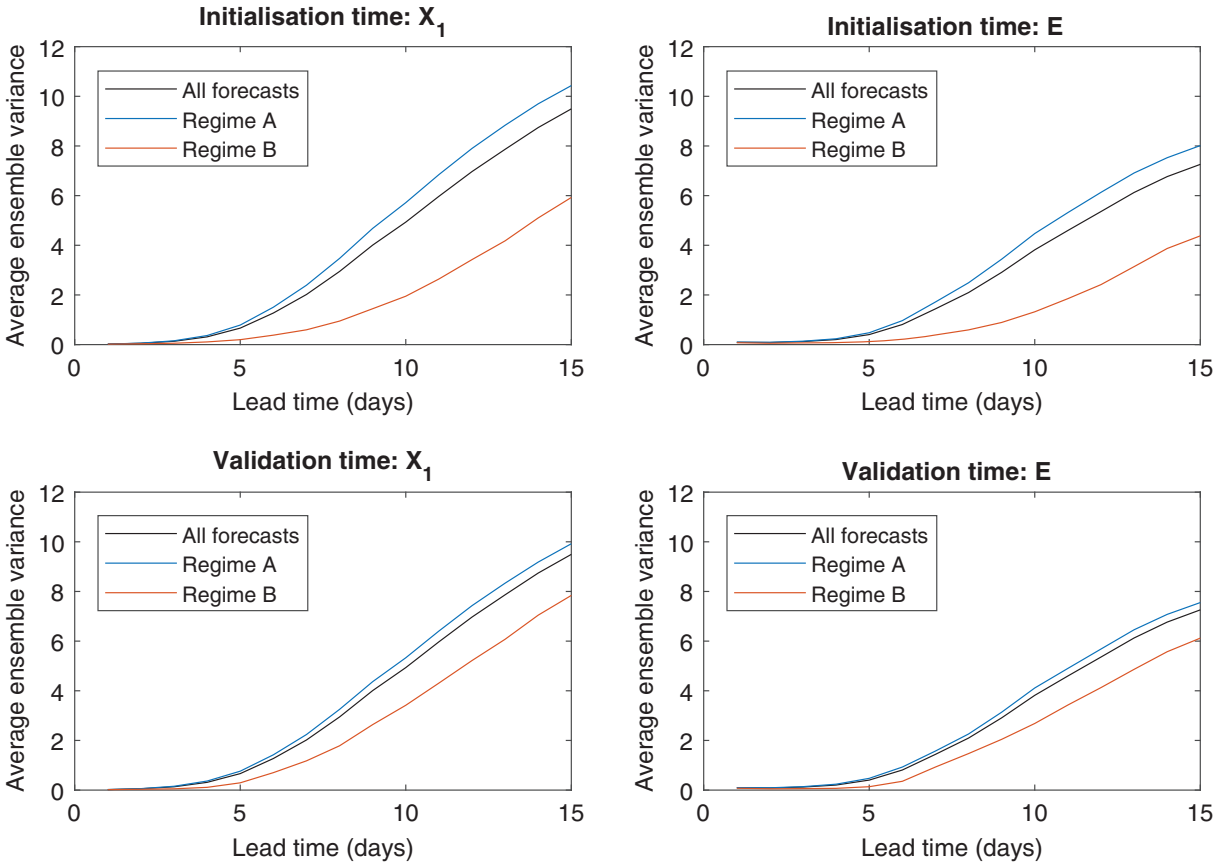


**FIGURE 4** Average ensemble variance for forecasts for $X_1$ (left) and $E$ (right), when initialised in each regime (top) and for those in each regime at the validation time (bottom)

largely overdispersed as a result of post-processing, indicating the forecasts are not calibrated conditional on the regime. Estimating a new set of parameters for forecasts initialised in regime B, as in RDNGR-init, helps to reduce the underconfidence of these forecasts.

Table 3 presents parameter estimates at a lead time of 7 days for both NGR and BMA, and all regime-dependent extensions. Regardless of the time at which the regime of the forecast is defined, the parameters, and in particular the variance coefficients, are noticeably different for the two regimes. For BMA, the parameter controlling the variance decreases dramatically when forecasting an event in regime B, supporting the belief that the system is more predictable in this regime. The regime A parameters, on the other hand, are generally similar to those obtained via standard post-processing. This is not surprising given that the system spends 80% of its time in regime A (Table 2) and hence the vast majority of forecasts in the training data are defined to be in regime A.

The regime-dependent NGR methods appear to adjust the variance of their predictive distribution differently for the two regimes. The modest ensemble spread in regime B forecasts is augmented by a larger scaling factor $\delta$, whereas the variance of regime A forecasts is increased by a larger additive, or nudging, parameter $\gamma$, thus implying the presence of a stronger spread-skill relationship in regime B. There are also slight differences between the parameters dictating the forecast mean. Reinforcing the results in Figure 4, such a pronounced difference in the variance parameters supports our theory that the forecast–observation relationship changes depending on the system's regime.

There do not appear to be large discrepancies between the regime-dependent approaches and it is difficult to deduce the time at which the forecast–observation relationships are most varied. The average ensemble variance displayed in Figure 4 is more contrasting when the regime is defined at the initialisation time yet the $\sigma_r^2$ estimates in Table 3 are more diverse for RDBMA-val than RDBMA-init, suggesting the validation time may produce slightly more heterogeneous relationships.

Having seen how the models are behaving in the different regimes, attention is turned to formally assessing the forecasts. Figure 6 exhibits the CRPS against lead time for the raw ensembles and for NGR, RDNGR-init, BMA and RDBMA-init forecasts, along with the breakdown of those defined to be in regime A and B at initialisation time. The scores are much lower for forecasts initialised in regime B than they are for regime A but since only 20% of forecasts are in regime B, the score calculated across all forecasts is more similar to that for forecasts in regime A. The post-processed forecasts unsurprisingly yield scores much lower than those for the raw ensemble forecasts and the improvements gained

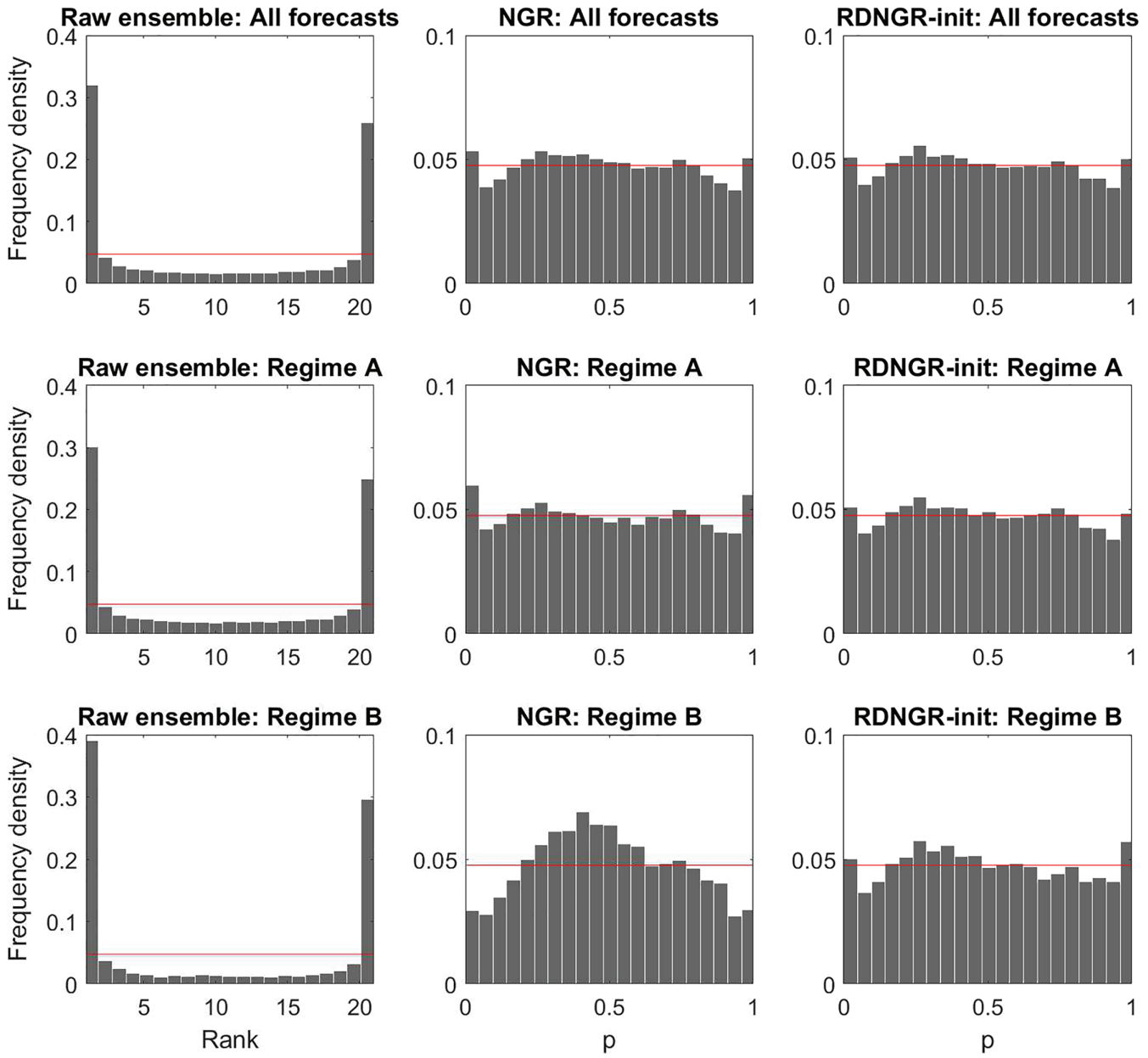


**FIGURE 5** Rank histograms for the raw ensemble forecasts and PIT histograms for NGR and RDNGR-init forecasts at a lead time of 7 days. Histograms are displayed for forecasts in each regime at initialisation time. The red line displays perfect uniformity and can hence be used as a comparison

from regime-dependent post-processing are noticeable in regime B but appear negligible for forecasts initialised in regime A, rendering the overall improvement relatively unpronounced. The CRPS for all methods at a lead time of 7 days is displayed in Table 4.

Equivalently, using the regime at the initialisation time allows the breakdown of skill-scores into regimes A and B. Figure 7 further reinforces what has already been seen: regime B forecasts improve by as much as 6% upon standard post-processing, while those initialised in regime A experience little improvement and even become marginally worse in cases. Regime B forecasts are thus responsible for the majority of improvement but the dominance of regime A means the relatively large improvements seen in regime B forecasts account for only 20% of the total improvement. Therefore, the maximum overall percentage improvement is little over 1%.

## 5.2 | Forecasting $E$

Figure 8 displays the evolution of BMA and RDBMA-init parameters over forecast lead time, when $E$ is the predictand. The variance coefficients exhibit similar behaviour to before, with $\sigma^2$ significantly lower for regime B forecasts than regime A forecasts.

However, as seen in Figure 2, the location of the distribution of observed values of $E$ in regime A is different to those in

**TABLE 3** Post-processing parameters for NGR and BMA, and for both of the regime-dependent extensions at a lead time of 7 days, when forecasting $X_1$

| $X_1$ | | $\alpha_r$ | $\beta_r$ | $\gamma_r$ | $\delta_r$ |
|---|---|---|---|---|---|
| NGR | | 0.301 | 0.884 | 4.476 | 2.713 |
| RDNGR-init | r = A | 0.411 | 0.855 | 6.028 | 2.202 |
| | r = B | −0.004 | 0.966 | 1.510 | 6.639 |
| RDNGR-val | r = A | 0.437 | 0.857 | 6.071 | 2.232 |
| | r = B | −0.140 | 0.976 | 1.212 | 4.414 |
| $X_1$ | | $\alpha_r$ | $\beta_r$ | $\sigma_r^2$ | |
| BMA | | 0.450 | 0.861 | 8.260 | |
| RDBMA-init | r = A | 0.569 | 0.831 | 9.499 | |
| | r = B | 0.091 | 0.961 | 4.065 | |
| RDBMA-val | r = A | 0.603 | 0.829 | 9.583 | |
| | r = B | −0.106 | 0.985 | 2.831 | |

regime B, which is not the case for verifications of $X_1$. There are now much larger distinctions in the location parameters, $\alpha$ and $\beta$, between the regimes, indicating the NWP model exhibits both spread and location biases that vary with the regime.

As a result, much larger improvements are gained from regime-dependent post-processing, as can be seen from the scores displayed in Figure 9. The scores for the raw ensembles are slightly lower than for forecasts of $X_1$. Nonetheless, the scores for RDNGR-val and RDBMA-val are considerably better than those for NGR and BMA respectively, particularly in regime B. This improvement is also maintained for forecasts at longer lead times.

Initially it was believed that RDBMA-val would have a slight advantage over its NGR counterpart since it post-processes each ensemble member separately, not compressing all the information into a single weight. NGR appears to yield more skilful forecasts than BMA overall but Figure 9 suggests the improvements are very similar for the two methods. When using the regime at the initialisation time, if the verification scores were smaller for BMA when the system resided in one regime but smaller for NGR when in the other, then it would be possible to calibrate subsets of forecasts using separate post-processing methods depending on the regime. i.e. apply NGR to all forecasts in regime A and BMA to all forecasts in regime B, for example.

The corresponding skill-scores are displayed in Figure 10. When the regime is defined at the validation time, forecasts in regime B can improve by almost as much as 20% on NGR and BMA forecasts, with overall improvements close to 7% at lead times between 6 and 9 days.

Given that regime A dominates the upper tail of the response distribution of $E$ (Figure 2) and regime B the lower,

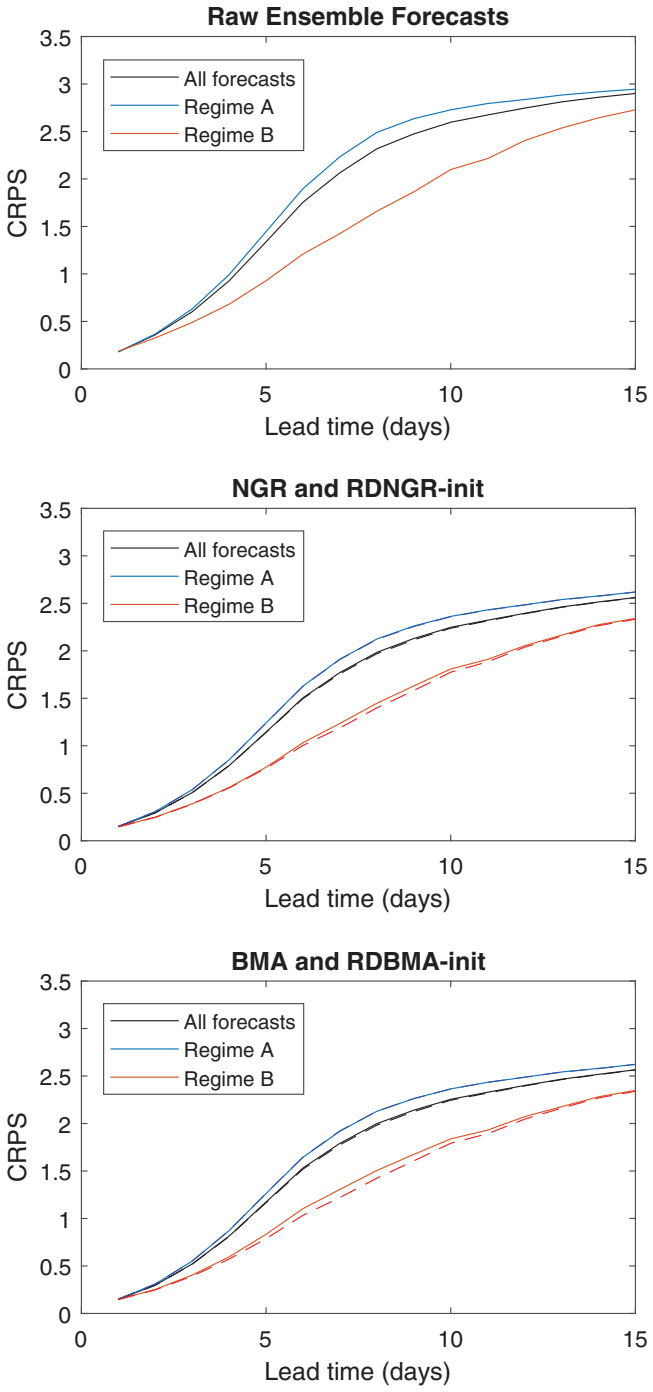


**FIGURE 6** CRPS for the raw ensemble forecasts of $X_1$ and for NGR and BMA (solid), and RDNGR-init and RDBMA-init (dashed) against lead time when the forecast is initialised in each regime

we might also expect regime-dependent post-processing to produce more informative predictions of extreme weather events. The Brier score, or mean squared error of a probability forecast for a binary response (Brier, 1950), can be used to assess the probability of the response falling above or below some threshold of the data.

Table 5 displays the Brier score, at lead times of 3, 5 and 10 days, for the predicted probability of the verification falling below the first percentile of all observations in the

**TABLE 4** CRPS for all forecasts of $X_1$ and the breakdown between those identified to be in regime A and regime B at initialisation time

| $X_1$ | Total | Regime A | Regime B |
|---|---|---|---|
| Raw | 2.072 (0.007) | 2.241 (0.008) | 1.429 (0.011) |
| NGR | 1.779 (0.005) | 1.921 (0.006) | 1.238 (0.008) |
| RDNGR-init | 1.768 (0.005) | 1.916 (0.006) | 1.202 (0.009) |
| RDNGR-val | 1.766 (0.005) | 1.914 (0.006) | 1.202 (0.008) |
| BMA | 1.796 (0.005) | 1.926 (0.006) | 1.301 (0.008) |
| RDBMA-init | 1.782 (0.005) | 1.927 (0.006) | 1.227 (0.009) |
| RDBMA-val | 1.779 (0.005) | 1.922 (0.006) | 1.232 (0.008) |

The scores are shown for the raw ensembles and for NGR, BMA post-processed forecasts, and all regime-dependent extensions, at a lead time 7 days. The corresponding standard errors are displayed in brackets next to the score.

training data. This is hence a measure of the forecasts' performance when predicting the occurrence of extremely low values of $E$. Again, when the regime is defined at the validation time regime-dependent statistical post-processing noticeably improves upon current post-processing approaches. Since the marginal distribution of $X_1$ varies less between the regimes, similar forecasts of extremely low values of $X_1$ exhibit less improvement, comparable to results seen for all forecasts in Figures 6 and 7.

## 6 | DISCUSSION AND CONCLUSIONS

This article acknowledges that the inability to distinguish between distinct relationships linking the NWP model and the atmosphere is a potential weakness of statistical techniques of calibrating ensemble forecasts. In particular, it is proposed that under certain circumstances the relationship between the model and atmosphere changes, and if such circumstances are identified then post-processing forecasts conditional on this extra information could yield more informative prognoses.

Although the methodology presented here extends to other appropriate and justifiable conditions, past literature suggests that the atmospheric circulation and, in particular, weather regimes are such circumstances. Section 2 discusses the relative merits of proposed ways of dealing with these new data. The continuous flow of the atmosphere can be represented by a comparatively small number of regime states and distinct post-processing parameters or methods can be used to calibrate forecasts for each regime separately.

This would suggest that different relationships arise from different subsets of the training data. The associated methods involve a simple division of the training data into relevant subsets from which separate parameters can be estimated. Although this may cause a problem if very few data are

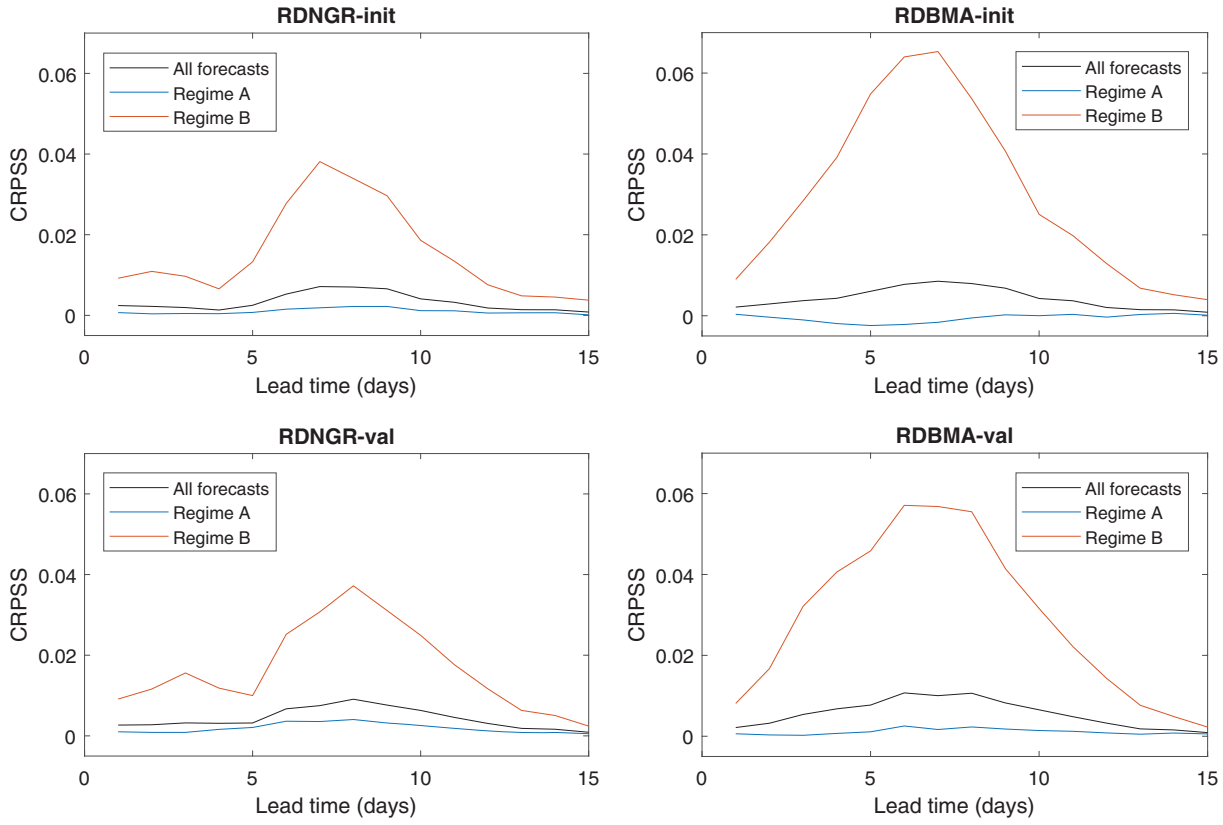


**FIGURE 7** CRPSS against lead time for both regime-dependent NGR and both regime-dependent BMA approaches using NGR and BMA, respectively, as a reference forecast when predicting $X_1$

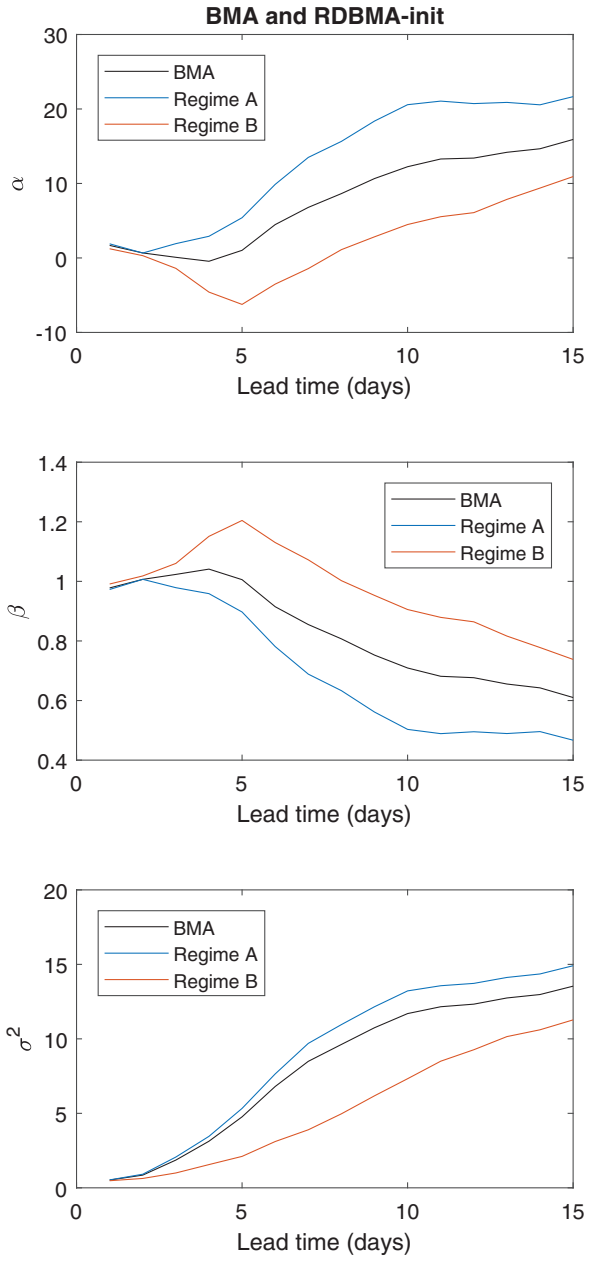


**FIGURE 8** Parameters for BMA forecasts of $E$ against lead time. RDBMA-init coefficients are also shown when the forecast is initialised in each regime

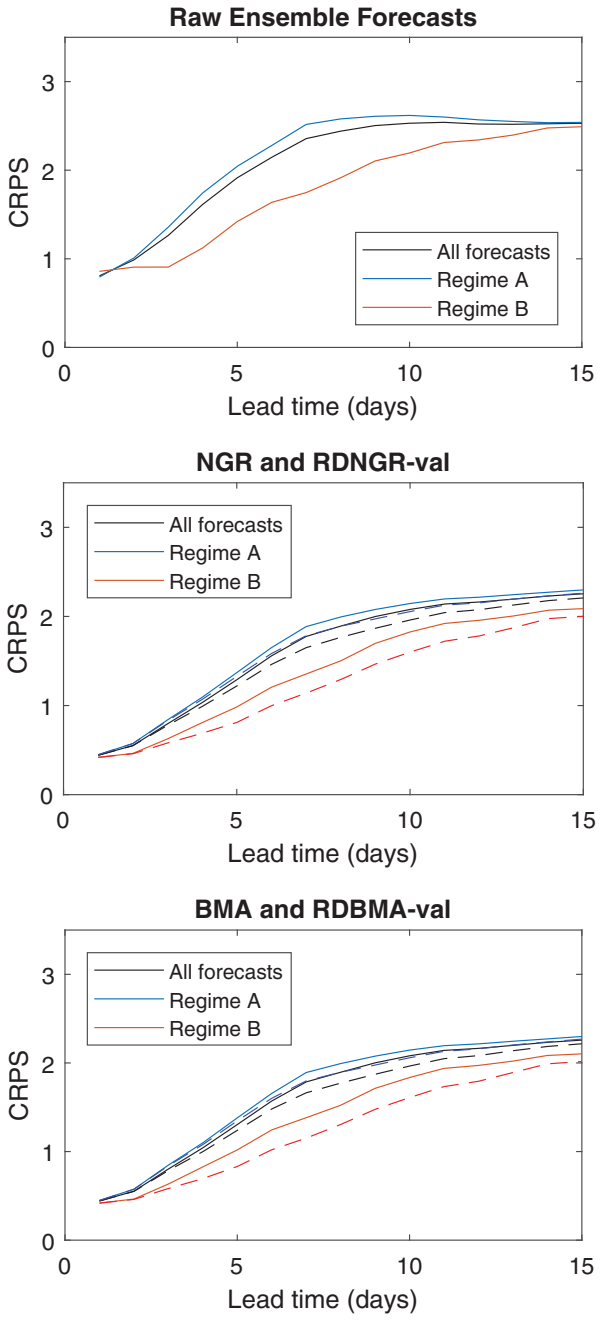


**FIGURE 9** CRPS for the raw ensemble forecasts of $E$ and for NGR and BMA (solid), and RDNGR-val and RDBMA-val (dashed) against lead time when the forecast is initialised in each regime

available for one of the subsets, defining regimes to exhibit persistence and recurrence renders this unlikely.

These subsets are constructed by defining the regime at the forecast's initialisation time and using observations to estimate the concurrent state of the atmosphere. This also allows forecasts in each regime to be evaluated separately. Establishing the regime at a time different to that at which the forecasts are being verified may not yield optimal improvements. It would be possible to use the regime at the forecast's validation time instead.

However, when issuing a new forecast, the regime at the validation time is not known and thus it is more appropriate to account for uncertainty in the regime using the state of the atmosphere predicted by the different ensemble members. There are a number of ways to utilize these regime predictions and they are used here to calculate weights for a mixture model forecast.

In this setting, a forecast–observation pair in the training data cannot be attributed to exactly one regime and hence rather than stratifying the data into MECE training subsets, all coefficients should be estimated simultaneously.

Defining the regimes at the validation time yields slightly larger improvements than when the initialisation time is used,

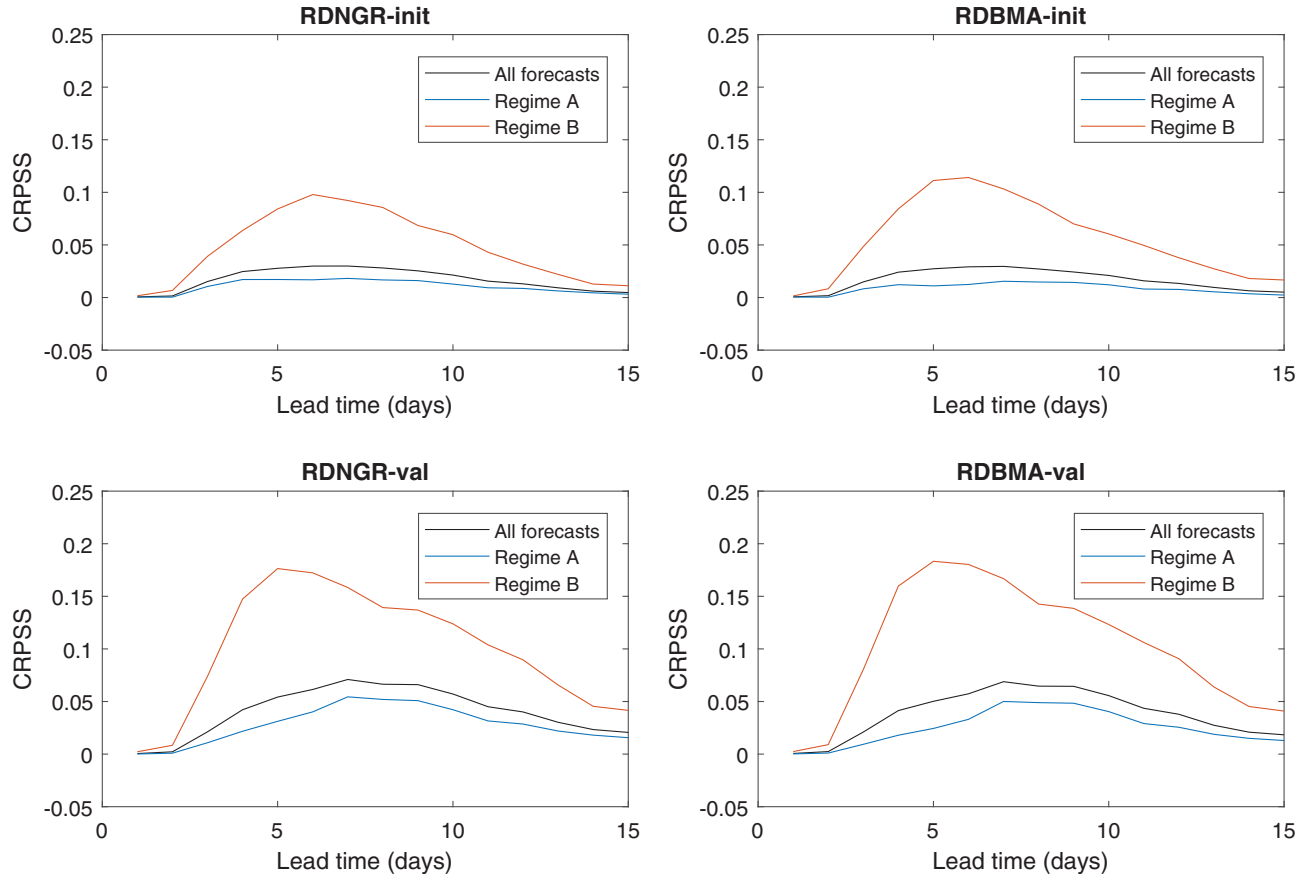


**FIGURE 10** CRPSS against lead time for both regime-dependent NGR and both regime-dependent BMA approaches using NGR and BMA, respectively, as a reference forecast when predicting $E$

**TABLE 5** Brier score for forecasts of the occurrence of extremely low values of $E$, for NGR, BMA and both regime-dependent extensions

| $E$ | 3 days | 5 days | 10 days |
|---|---|---|---|
| Raw | 4.59 (0.23) | 7.14 (0.32) | 12.25 (0.42) |
| NGR | 4.08 (0.20) | 6.52 (0.28) | 11.44 (0.44) |
| NGR-init | 3.46 (0.18) | 5.79 (0.26) | 11.34 (0.44) |
| NGR-val | 3.30 (0.17) | 4.95 (0.22) | 10.94 (0.40) |
| BMA | 4.09 (0.20) | 6.52 (0.27) | 11.42 (0.44) |
| RDBMA-init | 3.49 (0.18) | 5.61 (0.24) | 11.34 (0.44) |
| RDBMA-val | 3.31 (0.18) | 4.91 (0.22) | 10.92 (0.40) |

Extremely low values correspond to values below 29.1, the first percentile of the observations. Scores are shown at lead times of 3, 5 and 10 days, with the associated standard errors in brackets alongside. All values have been scaled by $10^3$.

**TABLE 6** Average time taken in seconds to estimate parameters at each forecast lead time for the different post-processing approaches, as implemented in MATLAB

| | $X_1$ | $E$ |
|---|---|---|
| NGR | 0.09 | 0.12 |
| RDNGR-init | 0.12 | 0.16 |
| RDNGR-val | 2.33 | 4.27 |
| BMA | 0.67 | 0.68 |
| RDBMA-init | 0.71 | 0.68 |
| RDBMA-val | 6.65 | 7.17 |

although estimating all regime-dependent parameters simultaneously can be significantly more computationally demanding than estimating BMA and NGR coefficients. On the other hand, despite the statistical models being more elaborate, implementing RDBMA-init and RDNGR-init was no more computationally expensive than the standard post-processing approaches in this study. The computational times for the different methods are shown in Table 6. Given that post-processing is typically done off-line, after integrating the forecast model, these regime-dependent approaches should not be prohibitively expensive.

These methods are trialled in the Lorenz (1996) system, a highly idealized model of the atmosphere involving only nonlinear advection, internal dissipation, external forcing and interactions between small- and large-scale variables. This system favours two states: regime A and regime B.

The results were compared for forecasts of two different variables, $X_1$ and $E$. These were chosen as predictands since the distribution of $X_1$ does not change much in the two regimes, whereas the opposite is true for $E$, yet in both cases the ensemble variance of forecasts in regime B is, on average, noticeably smaller than in regime A. Therefore, in both cases regime-dependent post-processing would be expected to calibrate forecasts differently in the regimes, and hence improve upon current post-processing approaches.

The fact that large improvements are seen for forecasts of $E$ but not for $X_1$ raises two questions. Firstly, why are improvements restricted when the response distribution does not change between the regimes, despite clear differences in the behaviour of the forecasts? Or alternatively, why does a varying response distribution contribute so much to improvements in forecast performance? And secondly, given the results presented here, how could regime-dependent post-processing be implemented in operational forecasting centres in the hope of attaining better forecasts of atmospheric variables?

Operational forecasters often suffer from a lack of historical data available, so persuading them to stratify these data further would be challenging. Furthermore, it has become common to use a sliding training window to estimate parameters. These windows consist of forecast–observation pairs from a relatively short number of days directly preceding the time of forecasting. The choice of the length of this window is a compromise between using enough data from which reliable parameter estimates can be obtained and using a length that is small enough for the training window to reflect the seasonality and recent behaviour of the weather.

The regime-dependent approaches estimate more parameters and hence require larger amounts of training data in order to attain reliable parameter estimates. Methods that can account for parameter uncertainty in the post-processing models (Siegert *et al.*, 2016) or augment the training data (Hamill *et al.*, 2017) are thus particularly desirable in the regime paradigm. An excessively large amount of training data was used in this simulation study to remove the necessity of such methods, although smaller archives of data drew the same conclusions.

It could be argued that knowing how the model behaves in different regimes is more valuable when estimating model coefficients than knowing how forecasts behaved more recently in potentially very different atmospheric conditions. For example, if the atmosphere resides in an anticyclonic regime then the model biases will likely be similar to occasions in previous years when this pattern has occurred, rather than to the errors, say, 20 days prior to forecasting when a different regime was present.

The method may thus be better suited to retrospective forecasting (reforecasting) approaches, that run current operational NWP models from historical analyses to generate a large number of hindcasts (Hamill *et al.*, 2004). This augmented dataset could then be used to estimate parameters. Although this could initially be computationally expensive, it reduces the need to estimate new post-processing parameters daily, and hence computational resources could be allocated to increasing the model resolution or complexity.

The results here suggest that if regimes can be identified such that the marginal distribution of the response changes, then regime-dependent post-processing can significantly improve weather forecasts. Moreover, if severe weather events occur more frequently in some regimes than others, such as extreme temperatures during prolonged blocking episodes, then incorporating this regime-dependency when calibrating forecasts could lead to refined predictions of these extreme events.

Results have only been presented for an NWP model that shows markedly different regime-like behaviour to that of the system it is modelling. The primary goal of Christensen *et al.* (2015) was to study the effects that stochastic parametrizations have on capturing the regime structure of the Lorenz (1996) system. The result was that the introduction of a red-noise stochastic parameter to the deterministic NWP model (Equation 4) provides a good estimation of the regimes. We repeated this study using the additive red-noise model used in Christensen *et al.* (2015), rather than a deterministic model, but chose to emphasize the method and its associated challenges rather than the characteristics of the model, and as such have not included these results.

It was found that similar patterns emerged to those identified here, but the improvements were slightly more pronounced using the deterministic model; the method was better at correcting poor forecasts than improving the higher-quality model, contradicting the idea that the method is reliant on the NWP model displaying similar regime-like characteristics to the true system. This behaviour is intuitive for atmospheric data; the circulation dictates the weather so the distribution of the observations would be expected to vary between regimes, and hence we would anticipate more improvement if the NWP model output did not do the same.

## ACKNOWLEDGEMENTS

Sam Allen was supported during this work by a NERC Industrial CASE studentship under grant reference NE/N008693/1. The authors thank Gavin Evans and Piers Buchanan of the UK Met Office for their helpful discussions and suggestions, particularly relating to the method's implementation in operational forecasting centres. This work has also benefitted from the comments of two anonymous reviewers, for which we are very grateful.

## ORCID

*Sam Allen* https://orcid.org/0000-0003-1971-8277

*Christopher A. T. Ferro* https://orcid.org/0000-0002-9830-9270
*Frank Kwasniok* https://orcid.org/0000-0003-1421-4010

## REFERENCES

Anderson, J.L. (1996) A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, 9, 1518–1530.

Baran, S. and Lerch, S. (2015) Log-normal distribution based Ensemble Model Output Statistics models for probabilistic wind-speed forecasting. *Quarterly Journal of the Royal Meteorological Society*, 141, 2289–2299.

Baran, S. and Lerch, S. (2016) Mixture EMOS model for calibrating ensemble forecasts of wind speed. *Environmetrics*, 27, 116–130.

Brier, G.W. (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1–3.

Cheng, X. and Wallace, J.M. (1993) Cluster analysis of the Northern Hemisphere wintertime 500-hpa height field: spatial patterns. *Journal of the Atmospheric Sciences*, 50, 2674–2696.

Christensen, H., Moroz, I. and Palmer, T.N. (2015) Simulating weather regimes: impact of stochastic and perturbed parameter schemes in a simple atmospheric model. *Climate Dynamics*, 44, 2195–2214.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39, 1–22.

Dole, R.M. and Gordon, N.D. (1983) Persistent anomalies of the extratropical Northern Hemisphere wintertime circulation: geographical distribution and regional persistence characteristics. *Monthly Weather Review*, 111, 1567–1586.

Ferranti, L., Corti, S. and Janousek, M. (2015) Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector. *Quarterly Journal of the Royal Meteorological Society*, 141, 916–924.

Fraley, C., Raftery, A.E. and Gneiting, T. (2010) Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Monthly Weather Review*, 138, 190–202.

Franzke, C., Woollings, T. and Martius, O. (2011) Persistent circulation regimes and preferred regime transitions in the North Atlantic. *Journal of the Atmospheric Sciences*, 68, 2809–2825.

Gneiting, T., Larson, K., Westrick, K., Genton, M.G. and Aldrich, E. (2006) Calibrated probabilistic forecasting at the stateline wind energy center: the regime-switching space–time method. *Journal of the American Statistical Association*, 101, 968–979.

Gneiting, T. and Raftery, A.E. (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102, 359–378.

Gneiting, T., Raftery, A.E., Westveld, A.H., III and Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133, 1098–1118.

Grimit, E.P., Gneiting, T., Berrocal, V. and Johnson, N.A. (2006) The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society*, 132, 2925–2942.

Hamill, T.M. (2001) Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129, 550–560.

Hamill, T.M. and Colucci, S.J. (1997) Verification of Eta–RSM short-range ensemble forecasts. *Monthly Weather Review*, 125, 1312–1327.

Hamill, T.M., Engle, E., Myrick, D., Peroutka, M., Finan, C. and Scheuerer, M. (2017) The US National Blend of Models for statistical postprocessing of probability of precipitation and deterministic precipitation amount. *Monthly Weather Review*, 145, 3441–3463.

Hamill, T.M., Whitaker, J.S. and Wei, X. (2004) Ensemble reforecasting: improving medium-range forecast skill using retrospective forecasts. *Monthly Weather Review*, 132, 1434–1447.

Hannachi, A., Straus, D.M., Franzke, C.L., Corti, S. and Woollings, T. (2017) Low-frequency nonlinearity and regime behavior in the Northern Hemisphere extratropical atmosphere. *Reviews of Geophysics*, 55, 199–234.

Horel, J.D. (1985) Persistence of the 500 mb height field during Northern Hemisphere winter. *Monthly Weather Review*, 113, 2030–2042.

Kimoto, M. and Ghil, M. (1993) Multiple flow regimes in the Northern Hemisphere winter. Part I: Methodology and hemispheric regimes. *Journal of the Atmospheric Sciences*, 50, 2625–2644.

Kondrashov, D., Ide, K. and Ghil, M. (2004) Weather regimes and preferred transition paths in a three-level quasigeostrophic model. *Journal of the Atmospheric Sciences*, 61, 568–587.

Kwasniok, F. (2012) Data-based stochastic subgrid-scale parametrization: an approach using cluster-weighted modelling. *Philosophical Transactions of the Royal Society*, 370, 1061–1086.

Leith, C. (1974) Theoretical skill of Monte Carlo forecasts. *Monthly Weather Review*, 102, 409–418.

Lerch, S. and Thorarinsdottir, T.L. (2013) Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A*, 65, 21206.

Lorenz, E.N. (1996) Predictability: a problem partly solved. In: *Proc. Seminar on Predictability*, Vol. 1. Reading, UK: ECMWF, 4–8 September 1995, pp. 1–18.

Majda, A.J., Franzke, C.L., Fischer, A. and Crommelin, D.T. (2006) Distinct metastable atmospheric regimes despite nearly Gaussian statistics: a paradigm model. *Proceedings of the National Academy of Sciences*, 103, 8309–8314.

Messner, J.W., Mayr, G.J. and Zeileis, A. (2017) Nonhomogeneous boosting for predictor selection in ensemble postprocessing. *Monthly Weather Review*, 145, 137–147.

Neal, R., Fereday, D., Crocker, R. and Comer, R.E. (2016) A flexible approach to defining weather patterns and their application in weather forecasting over Europe. *Meteorological Applications*, 23, 389–400.

Raftery, A.E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005) Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133, 1155–1174.

Robertson, A.W. and Ghil, M. (1999) Large-scale weather regimes and local climate over the western United States. *Journal of Climate*, 12, 1796–1813.

Roulston, M.S. and Smith, L.A. (2003) Combining dynamical and statistical ensembles. *Tellus A*, 55, 16–30.

Siegert, S., Sansom, P.G. and Williams, R.M. (2016) Parameter uncertainty in forecast recalibration. *Quarterly Journal of the Royal Meteorological Society*, 142, 1213–1221.

Smyth, P., Ide, K. and Ghil, M. (1999) Multiple regimes in Northern Hemisphere height fields via mixture model clustering. *Journal of the Atmospheric Sciences*, 56, 3704–3723.

Stephenson, D., Coelho, C., Doblas-Reyes, F. and Balmaseda, M. (2005) Forecast assimilation: a unified framework for the combination of multi-model weather and climate predictions. *Tellus A*, 57, 253–264.

Talagrand, O. (1997) Evaluation of probabilistic prediction systems. In: *Proc. ECMWF Workshop on Predictability*. Reading, UK: ECMWF, 20–22 October 1997, pp. 1–18.

Van Schaeybroeck, B. and Vannitsem, S. (2015) Ensemble post-processing using member-by-member approaches: theoretical aspects. *Quarterly Journal of the Royal Meteorological Society*, 141, 807–818.

Wallace, J.M. and Gutzler, D.S. (1981) Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Monthly Weather Review*, 109, 784–812.

Wilks, D.S. (2005) Effects of stochastic parametrizations in the Lorenz '96 system. *Quarterly Journal of the Royal Meteorological Society*, 131, 389–407.

Wilks, D.S. (2006) Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteorological Applications*, 13, 243–256.

Wilks, D.S. (2019) *Statistical Methods in the Atmospheric Sciences*. Amsterdam: Elsevier.

Williams, R.M., Ferro, C.A.T. and Kwasniok, F. (2014) A comparison of ensemble post-processing methods for extreme events. *Quarterly Journal of the Royal Meteorological Society*, 140, 1112–1120.

**How to cite this article:** Allen S, Ferro CAT, Kwasniok F. Regime-dependent statistical post-processing of ensemble forecasts. *Q J R Meteorol Soc*. 2019;145:3535–3552. https://doi.org/10.1002/qj.3638